Supplementary: Simultaneous discovery of cancer subtypes and subtype features by molecular data integration

Thanh Le Van [1], Matthijs van Leeuwen [2], Ana Carolina Fierro [3], Dries De Maeyer [4], Jimmy Van den Eynden [5], Lieven Verbeke [3], Luc De Raedt [1], Kathleen Marchal [3,4,6,7], Siegfried Nijssen [1,2]

[1]Department of Computer Science, KULeuven, Belgium,
[2]Leiden Institute for Advanced Computer Science, Universiteit Leiden, The Netherlands,
[3]Department of Information Technology, iMinds, Ghent University, Belgium,
[4]Bioinformatics Institute Ghent, Technologiepark 927, 9052 Gent, Belgium,
[5]Department of Medical Biochemisty and Cell Biology, Institute of Biomedicine, University of Gothenburg, Sweden,
[6]Department of Plant Biotechnology and Bioinformatics, Ghent University, Belgium,
[7]Department of Genetics, University of Pretoria, Hatfield Campus, Pretoria 0028, South Africa.

1 Diffusion threshold selection

Our SRF algorithm transforms the Boolean mutation matrix into a rank matrix by applying a diffusion model [1], [2] for each mutation profile of the tumor samples on a protein–protein interaction network. Intuitively, the diffusion model links mutated genes to its neighbours with respect to the interaction network structure. The strength of these links is governed by a tuning parameter called α. Hofree et. al., [2] found the optimal value of α is network dependent. In the case of the STRING [3] network, they also discovered that the optimal value for α is 0.7. As we used the STRING network, which was post-processed by Hofree et. al., [2], we set the α threshold to 0.7 for all of the experiments presented in this paper.

2 SRF parameter selection

Our algorithm has five parameters: θ_1 and θ_2 specify the minimal threshold on respectively the diffusion and expression ranks for a gene/sample to be included in a ranked factor; β specifies the importance of the mutation relative to the expression data; k specifies the number of ranked factors; μ specifies the number of patients in which a mutation should be present in order to be included in a factor. μ was set to 2 for all the experiments in this paper.

For each simulated dataset, we performed a parameter sweep using combinations of the following parameter settings: $\theta_1 \in \{82\%, 85\%, 90\%\}, \theta_2 \in \{65\%, 70\%\}$,

and $\beta \in \{2, 18, 35, 50\}$. As we knew the ground truth of these datasets, we evaluated F1 scores and chose the parameter setting that resulted in the mean highest average score over all simulated datasets. This resulted in the following choice for the parameters: $\theta_1 = 82\%, \theta_2 = 65\%, \beta = 18$. Note that with this choice of β, the mutation component is implicitly given two times $(max(D)/(\beta * max(E))$, where $max(D) = 12232$ and $max(E) = 350$) the weight of the expression component in the optimisation problem.

Given that our artificial data has similar properties as the TCGA data, for the TCGA data we used the same parameter settings for θ_2 and β as for the artificial data, i.e., $\theta_2 = 65\%$ and $\beta = 18$; also in an earlier study [4], on expression data only, it was demonstrated that $\theta_2 = 65\%$ is a good choice. We considered alternative settings for θ_1, with $\theta_1 \in \{70\%, 72\%, \ldots, 90\%, 92\%\}$. Our motivation is that we wished to end up with a number of mutations smaller than 40, which is the number of cancer genes of this disease found by [5]. Although the impact of the θ_1 parameter is small, we decided to use a parameter setting of $\theta_1 = 86\%$ to reduce the set of mutated genes. For k we considered values in the range $k \in \{5, \ldots, 14\}$. We observed that for $k > 8$ the results only change slowly and hence stopped at $k = 8$.

To validate whether our choice is reasonable for the TCGA dataset or not, we ran SRF with a selected number of combinations of the parameter thresholds, i.e., varying one parameter while fixing the others to the selected values described above. Then, we evaluated two scores: *coverage* and *error*. The coverage score is the percentage of the region in the reconstructed matrix of the factorisation that has non-zero values. The error score is the average number of cells in the covered region whose value is less than the user-defined threshold.

Figure 1 shows the behaviour of the algorithm when we varied the value of β (the relative importance threshold of mutation to expression) and fixed the other parameters to the selected values mentioned above ($\theta_1 = 86\%, \theta_2 = 65\%, k = 8$). The figure confirms that $\beta = 18$ was a good choice as from that point the mutation coverage levelled off and became reasonably small. At the same time, the mutation error per cell slowly increased since $\beta = 18$.

Figure 2 illustrates the performance of the algorithm when we varied θ_1 (the rank threshold for ranked diffusion) and fixed the other parameters to the selected values. The figure also confirms that $\theta_1 = 86\%$ was a good choice as the mutation coverage (and hence the number of mutations per subtype) became reasonably small since that point. At the same time, the expression coverage started increasing while the expression error per cells decreased. In other words, high quality of expression data were added into the solutions since $\theta_1 = 86\%$.

Figure 3 shows the behaviour of the algorithm when we varied θ_2 (the rank threshold for ranked expression) and fixed the others to the selected values. The figure shows $\theta_2 = 70\%$ could be a good choice as both mutation error and expression error decreased at that point. However, the mutation coverage increased and hence the number of mutations per subtype also increased. Because we wished to end up with a number of mutations smaller than 40, which is the

number of cancer genes of this disease found by Stephens et al. [5], $\theta_2 = 65\%$ was a good selection.

Figure 4 shows that the mutation coverage became stable since $k = 8$ while expression coverage slowly increased. Hence, we could stop at $k = 8$.

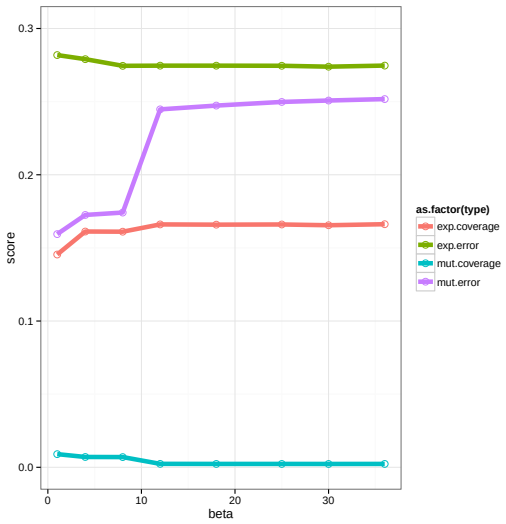

Fig. 1: Varying the value of β while fixing $\theta_1 = 86\%, \theta_2 = 65\%, k = 8$

3 Parameter selection for the benchmarked methods

With iCluster+ [6], we used the model selection algorithm provided by the software to obtain the optimal parameters. With SNF [7], we varied the α parameter in the range of $\{0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$ and chose the one that resulted in the highest average F1 score on the simulated data. With NBS [2], we used the default parameter settings. Note that the two scores were calculated for both of the two rank matrices.

References

1. Vanunu, O., Magger, O., Ruppin, E., Shlomi, T., Sharan, R.: Associating genes and protein complexes with disease via network propagation. PLoS computational biology **6**(1) (January 2010) e1000641
2. Hofree, M., Shen, J.P., Carter, H., Gross, A., Ideker, T.: Network-based stratification of tumor mutations. Nature methods **10**(11) (November 2013) 1108–15
3. Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., et al.: The string database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Research **39**(suppl 1) (2011) D561–D568

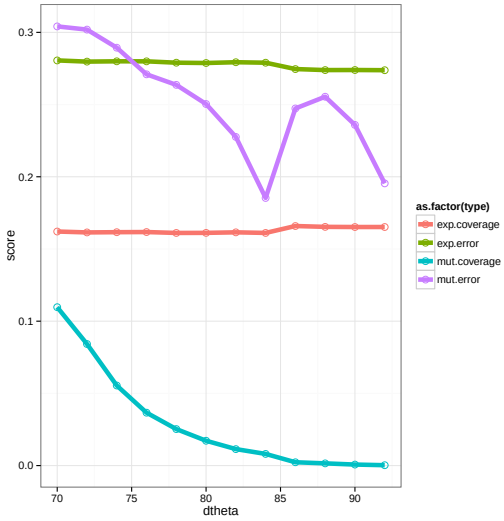

Fig. 2: Varying the value of θ_1 while fixing $\theta_2 = 65\%, \beta = 18, k = 8$

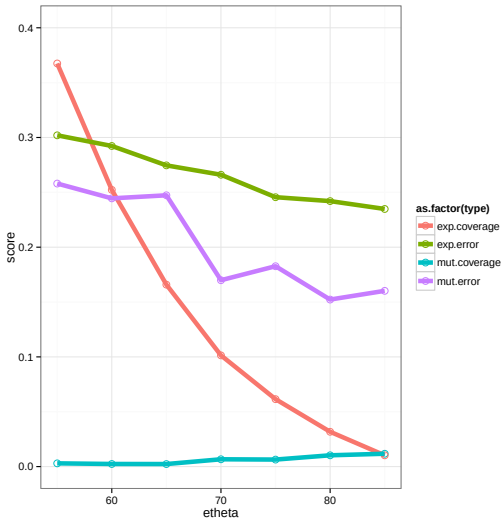

Fig. 3: Varying the value of θ_2 while fixing $\beta = 18, \theta_1 = 86\%, k = 8$

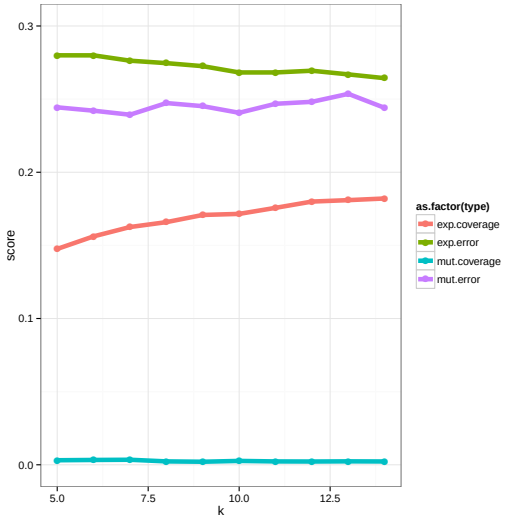

Fig. 4: Varying k while fixing $\beta = 18, \theta_1 = 86\%, \theta_2 = 65\%$

4. Le Van, T., van Leeuwen, M., Nijssen, S., Fierro, A.C., Marchal, K., De Raedt, L.: Ranked tiling. In: ECML PKDD 2014 (2). (2014) 98–113
5. Stephens, P.J., Tarpey, P.S., Davies, H., Van Loo, P., et al.: The landscape of cancer genes and mutational processes in breast cancer. Nature **486**(7403) (2012) 400–4
6. Mo, Q., Wang, S., Seshan, V.E., Olshen, et al.: Pattern discovery and cancer gene identification in integrated cancer genomic data. PNAS **110**(11) (March 2013) 4245–50
7. Wang, B., Mezlini, A.M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B., Goldenberg, A.: Similarity network fusion for aggregating data types on a genomic scale. Nature methods **11**(3) (2014) 333–7